

## ■ QUALITÄT UND RELEVANZ VON BILDUNGSBEZOGENEN SUCHERGEBNISSEN BEI DER SUCHE IM WEB

von Katrin Geist

### Inhalt

1. Einleitung – Bildungsbezogene Informationssuche
2. Evaluation von Websuchmaschinen
3. Testdesign dieser Studie
4. Auswertung der Relevanz- und Qualitätsurteile der Nutzer
5. Ausblick

**Zusammenfassung:** Websuchmaschinen haben sich sowohl im privaten als auch im professionellen Bereich zu den bedeutsamsten Recherchewerkzeugen der Gegenwart entwickelt. Sie durchsuchen die Weiten des Internets, um verfügbare Dokumente zu erfassen und dem Benutzer auf Anfrage präsentieren zu können. Mit den gefundenen Dokumenten beabsichtigen sie vielfältige Informationsbedürfnisse zu befriedigen. Ihr Erfolg basiert zu einem großen Teil auf der Fähigkeit, relevante Dokumente zu spezifischen Suchanfragen zu liefern. Um verschiedene Systeme miteinander vergleichen zu können, werden daher häufig Relevanzevaluationen durchgeführt.

Dieser Artikel beschreibt in Auszügen die Ergebnisse eines Retrievaltests, der Qualität und Relevanz informationsorientierter Suchanfragen zum Thema Bildung untersucht. Studentische Nutzer beurteilten dabei die Suchergebnisse von Google hinsichtlich ihrer Relevanz, Vertrauenswürdigkeit, Verständlichkeit und Aktualität. Die Untersuchung wurde im Rahmen meiner Masterarbeit durchgeführt, die mit dem VFI-Förderungspreis 2011 ausgezeichnet wurde, und war eingebunden in ein Forschungsprojekt an der Universität Hildesheim zu bildungsbezogener Informationssuche im Internet (BISIBS). Ich bedanke mich herzlich bei allen, die an der Entstehung dieser Arbeit beteiligt waren und bei der Preis-Kommission für die Auszeichnung.

**Schlagwörter:** Websuchmaschinen, Fachinformation, nutzergenerierte Inhalte, Relevanzevaluation, Retrievaleffektivität, Informationsqualität

## INFORMATION QUALITY AND RELEVANCE OF WEB SEARCH RESULTS FOR EDUCATION-RELATED QUERIES

**Abstract:** Web search machines have become the most significant tool for research in both the private and professional sphere, browsing the expanse of the internet in order

*to gather available documents and present them upon request to the user. They aim to satisfy a diverse array of information needs with the documents found, and their success depends largely upon their ability to deliver documents relevant to specific search queries. As such, evaluations of relevance are often carried out in order to compare different information retrieval systems.*

*This article describes some of the results of a retrieval test which analysed the quality and relevance of information-oriented search queries in the field of education. Student users judged Google organic search results in terms of their relevance, reliability, intelligibility and topicality. The research was carried out within the scope of a master thesis that was awarded the VFI Research Prize in 2011 and incorporated into a research project at the University of Hildesheim regarding education related information search on the internet (BISIBS). I thank all those involved in the development of this study and also the Prize Commission for the aforementioned award.*

**Keywords:** *web search, specialized information, user-generated content, retrieval test, results quality, retrieval effectiveness*

## **1. Einleitung**

Die Untersuchung bildungsbezogener Informationssuche richtet den Blick auf die Suche nach Internetquellen verschiedenster Art, die sich im weitesten Sinne mit Bildung beschäftigen (z.B. Bildungssystem, Bildungspolitik, Bildungsangebote). Dazu zählt auch die Suche nach Fachinformationen für Lehr- und Lernprozesse. Bisher finden sich kaum wissenschaftliche Untersuchungen zu bildungsbezogenem Informationsverhalten oder Erkenntnisse zur Ressourcenauswahl, zu Relevanzgesichtspunkten, Dokumentselektion oder der Reflexion des Suchprozesses.

Die Nutzer von Suchmaschinen suchen jedoch nach den unterschiedlichsten Themen, die ihr Leben betreffen. Dies gilt sowohl für professionelle als auch private Themen. Universalsuchmaschinen, wie beispielsweise Google, bieten weder eine semantische Eingrenzung auf fachliche Inhalte noch berücksichtigen sie fachspezifische Besonderheiten oder wissenschaftliche Standards. Dennoch werden sie häufig für fachbezogene Suchanfragen genutzt und liefern relevante Treffer.

Lewandowski kommt 2006 in einer Untersuchung von Suchanfragen deutscher Suchmaschinen zu dem Ergebnis, dass im Themenbereich Bildung und Geisteswissenschaften 48,4% der Suchanfragen deutschsprachiger Nutzer informationsorientierter, 35,5% navigationsorientierter und 16,1%

transaktionsorientierter Natur sind. Diese Unterscheidung der Anfragetypen geht auf Broder (2002) zurück, der die Auswirkung von verschiedenen Informationsbedürfnissen auf das Suchverhalten thematisiert. Navigationsorientierte Anfragen zielen darauf ab, ein Dokument wiederzufinden, von dessen Existenz der Nutzer weiß oder zumindest eine Vorstellung hat. Die Suche nach der Webseite einer bestimmten Institution (z.B. „DIPF“ – Deutsches Institut für Internationale Pädagogische Forschung) oder nach einer Person (z.B. „Annette Schavan“) sind Beispiele für derartige Informationsbedürfnisse. Mit informationsorientierten Anfragen möchte sich der Nutzer über ein Thema (z.B. Entwicklungspsychologie) informieren, daher werden in der Regel mehrere Dokumente benötigt. Transaktionsorientierte Anfragen finden mit der Absicht statt, nach dem Aufruf einer Webseite eine Interaktion durchzuführen. Transaktionen sind beispielsweise das Aufgeben einer Bestellung oder der Download einer Datei.

Diese Unterscheidung hat sich bei der Untersuchung des Suchverhaltens mit Suchmaschinen etabliert und findet häufig Anwendung. Von der Berücksichtigung der Suchanfragetypen erhofft man sich beispielsweise Verbesserungen der Retrievaleffektivität und Personalisierung der Suche. Dennoch lässt sich die Absicht hinter Suchanfragen nicht immer einwandfrei aus konkreten Suchanfragen erschließen. Eine allgemein formulierte Suchphrase wie „Legasthenietherapie“ kann navigationsorientiert, informationsorientiert oder transaktionsorientiert sein.

Die durchgeführte Untersuchung berücksichtigt ausschließlich informationsorientierte Suchanfragen. Somit steht in dieser Untersuchung das Bedürfnis nach hochwertiger Fachinformation der Realität im Web gegenüber. Dynamik und stete Veränderung der Dokumente sind allgegenwärtig. Ununterbrochen entstehen neue Inhalte und die Menge nutzergenerierter Inhalte übersteigt die der professionell erstellten Inhalte. Jede Minute werden beispielsweise 60 Stunden Videomaterial auf YouTube hochgeladen. In einem Monat sind dies mehr Videos als von den drei großen amerikanischen Sendern in 60 Jahren erstellt wurden<sup>1</sup>. Die Fotoplattform Flickr hat mit Hilfe der Nutzer über 7 Milliarden Bilder innerhalb der letzten 8 Jahre angehäuft<sup>2</sup>. Wie viele Blogs es allein in Deutschland gibt, lässt sich kaum beantworten und hängt von der Sichtweise des Autors ab. Zählen Twitter-Accounts, MySpace-Seiten oder Pinterest-Profile dazu? Einer Studie von NM Incite zufolge gibt es Ende 2011 181 Millionen Blogs weltweit, die mittels Blogging-Plattformen und Sozialen Netzwerken publizieren<sup>3</sup>. Baeza-Yates (2009) geht davon aus, dass nutzergenerierte Inhalte auch für Websuch-

maschinen relevant sind und stellt die These in den Raum, dass die Qualität der Inhalte von entscheidender Bedeutung ist. Es finden sich umfangreiche Forschungsarbeiten, die die Qualität nutzergenerierter Inhalte untersuchen (siehe hierzu Chai et al. 2009). Informationsqualität als Konzept bildet nahezu einen eigenen Forschungsbereich innerhalb der Informationswissenschaften, der danach strebt, die Beschaffenheit von Dokumenten zu erfassen. Die Auseinandersetzung mit Informationsqualität hat zu einer Vielzahl von Frameworks geführt, die versuchen, Informationsqualität messbar zu machen. Die Kontextabhängigkeit von Informationsqualität spielt dabei eine wichtige Rolle. Knight & Burn (2005) bieten einen guten Überblick. Die Analyse der Frameworks zeigt, dass sich gewisse Elemente in verschiedenen Forschungsarbeiten wiederfinden lassen. Dazu zählen beispielsweise Fehlerfreiheit, Kontinuität, Aktualität, Vollständigkeit, Zugänglichkeit, Objektivität und Relevanz.

Die Erfassung von Informationsqualität bzw. von relevanten Dimensionen ist in der Praxis eine Herausforderung, da die Beurteilung der Qualität eines Informationsangebots in der Regel in einer sehr kurzen Zeitspanne und kontinuierlich erfolgt (Rieh 2002). Hinzu kommt, dass je nach Informationsobjekt (z.B. kommerzielle Angebote oder private Webseiten), Situation, Wissenstand etc. unterschiedliche Dimensionen im Vordergrund stehen (Metzger 2007).

Die Verbindung von Informationsqualität und Relevanz mit verschiedenen Suchergebnistypen ist bisher kaum behandelt worden (Dickinger 2011, Lewandowski/Spree 2011). In der hier vorgestellten Untersuchung wird die Relevanzbeurteilung von Suchergebnissen um Urteile zur Informationsqualität ergänzt. Dazu werden beispielhaft drei Dimensionen ausgewählt: Vertrauenswürdigkeit, Verständlichkeit und Aktualität. Die Auswahl dieser Attribute wird durch ihre Bedeutung für die Relevanzbeurteilung begründet. Vertrauenswürdigkeit ist ein Attribut auf Ebene der intrinsischen Qualität der Informationsressourcen. Es ist ein zentrales Kriterium bei der Beurteilung von Informationen im Internet. Verständlichkeit fällt in den Bereich der Präsentationsqualität und thematisiert, inwiefern Information erfasst werden kann. Aktualität ergänzt auf der Ebene der kontextuellen Qualität die Relevanzurteile. Diese drei Faktoren wurden auch von Gammer et al. (2008) bei einem Retrievaltest zum Vergleich der Effektivität von algorithmischen Suchmaschinen, Social Bookmarking-Diensten und Frage-Antwort-Diensten verwendet.

## 2. Evaluation von Websuchmaschinen

Websuchmaschinen sind mit einer riesigen Menge heterogener Dokumente konfrontiert, die eine Vielzahl von Informationsbedürfnissen einer heterogenen Nutzerschaft bedienen sollen. Es ist daher von entscheidender Bedeutung, dass sie in der Lage sind, relevante Ergebnisse zu liefern. Relevanz ist seit jeher das zentrale Konzept des Information Retrieval. Die meisten Websuchmaschinen ordnen ihre Ergebnisse, um die relevantesten Dokumente zuerst zu präsentieren. Zur Evaluation von klassischen Information Retrieval Systemen wurden Retrievaltests entwickelt, die mit einer Testkollektion von Dokumenten, auf Informationsbedürfnissen basierenden Suchanfragen und der Relevanzeinstufung von Dokumenten arbeiten. Effektive Retrievalsysteme sollen relevante Dokumente liefern, während sie gleichzeitig nicht-relevante Dokumente zurückhalten. Websuchmaschinen sind jedoch mit einer deutlich komplexeren Situation konfrontiert. Daher schlagen Lewandowski & Höchstötter (2007) eine Qualitätsmessung vor, die über die Bewertung der gefundenen Dokumente hinausgeht und die Qualität des Indexes, der Suchfunktionen und die Nutzerfreundlichkeit ebenfalls berücksichtigt.

Für die vorliegende Untersuchung wird ein Retrievaltest konzipiert, der sich am Aufbau klassischer Retrievaltests orientiert und lediglich die Trefferqualität berücksichtigt. Dies ist unter anderem in der Effektivität der Testdurchführung begründet. Die Relevanzbeurteilung der Suchergebnisse wird jedoch um zusätzliche Qualitätsdimensionen ergänzt. Ein weiterer Untersuchungsschwerpunkt liegt auf möglichen Beziehungen zwischen der Qualität und der Relevanz verschiedener Suchergebnistypen. Professionalität und Fachbezug sind daher wichtige Aspekte bei der Beschreibung der Suchergebnisse. Die Ergebnisdokumente werden in drei Kategorien unterteilt: Fachinformation (FI), nutzergenerierte Inhalte (UGC) und professionell erstellte Inhalte (P). Den Ausgangspunkt der Zuordnung bildet die Absicht, die qualitative Beurteilung nutzergenerierter Inhalte mit der Beurteilung von Fachinformationen zu vergleichen.

## 3. Testdesign

In diesem Abschnitt werden kurz die methodische Konzeption und die praktische Umsetzung des Retrievaltests dargestellt. Anhand des thematisch begrenzten Untersuchungsbereichs bildungsbezogener Suchanfragen sollen die erhobenen Daten zeigen, dass es bedeutsame Unterschiede bei der Beurteilung der Ergebnistypen gibt.

## ***Suchanfragen und Informationsbedürfnisse***

Aus Gründen der Effektivität werden 40 Suchanfragen für den Retrievaltest ausgewählt. Diese entstammen zwei verschiedenen Quellen. Die eine Hälfte ist eine zufällige Auswahl aus den 500 häufigsten Suchphrasen bei externen Suchmaschineneinstiegen zu Seiten des Deutschen Bildungsservers (DBS)<sup>4</sup>. Die andere Hälfte wird mit Hilfe einer Befragung von 123 sich im Masterstudium befindenden Lehramtsstudierenden an der Universität Hildesheim gewonnen. Die Studierenden werden gebeten, jeweils zu einer der zufällig ausgewählten Suchanfragen des DBS ein Informationsbedürfnis zu formulieren und eine selbst vor Kurzem getätigte Suchanfrage zu beschreiben. Die Rekonstruktion der Informationsbedürfnisse wird anschließend in einem mehrstufigen Prozess durchgeführt. Die Angaben werden von zwei Jurorinnen diskutiert und manuell aufbereitet. Ziel ist es, die Angaben der Studierenden so wenig wie möglich zu verändern und dennoch zu gewährleisten, dass Spezifik und Umfang des Informationsbedürfnisses in einem ausgewogenen Verhältnis stehen. Der Korpus der Suchanfragen lässt sich als ausgeglichenes Sample von tatsächlich getätigten und sorgfältig konstruierten Anfragen typischer Nutzer beschreiben.

### ***Erschließung der Ergebnisse***

Die Suchanfragen werden alle am 30.03.2011 über den Internet-Browser Firefox (3.6.16) auf Google.de abgefragt. Die Sucheinstellung wird insofern angepasst, als je 20 Treffer pro Ergebnisseite sichtbar sind. Die Ergebnisseiten der Top-20 organischen Ergebnisse werden gespeichert und als Grafik (JPEG-Format, 100%, farbig) mit dem Firefox Plugin FireShot (Version 0.91) erfasst. Um eine Personalisierung von Google zu erschweren, wird vor jeder Abfrage die Verbindung zum Netzwerk getrennt, sodass eine neue IP-Adresse vergeben wird. Außerdem wird der gesamte Verlauf des Browsers (Suchverlauf, Browserverlauf, Downloadverlauf und Cookies) nach jeder Anfrage gelöscht. Die standortbasierte Anpassung (Hannover) ist nicht deaktiviert. Die Suchsprache ist Deutsch. Spezielle Dateiformate (z.B. PDF, DOC, PPT) werden separat gespeichert.

Nach der Erschließung der Ergebnisse werden diese manuell den Ergebniskategorien Fachinformation, professionelle Informationsangebote und nutzergenerierte Inhalte zugeordnet (siehe Tabelle 1). Für die Zuordnung zu den Ergebnistypen sind Professionalität und Fachlichkeit entscheidend. Daher werden bei der Typisierung an jedes Dokument zwei Fragen gestellt:

Ist der Informationsanbieter in einen professionellen Arbeitskontext eingebunden? Ist der Anbieter im Bildungssektor beschäftigt? Informationsanbieter können Organisationen, Unternehmen, Gruppen und Projekte oder Privatpersonen sein.

		Professionelle Arbeitsroutinen	
		nein	ja
Akteur mit Bildungsauftrag	nein	Nutzergenerierte Inhalte (UGC)	Professionell erstellte Inhalte (P)
	ja	-----	Fachinformation (FI)

Tabelle 1: Die Zuordnung der Ergebnisdokumente zu Ergebniskategorien

Im Vordergrund steht, wer für die Erstellung von Inhalten verantwortlich ist. Es lässt sich argumentieren, dass es eine vierte Gruppe von Anbietern gibt, die in dieser Auswahl nicht angemessen berücksichtigt wird: im Bildungssektor Beschäftigte (Experten), die in ihrer Freizeit als Privatperson Information im Internet veröffentlichen. Im vorliegenden Ergebniskorpus sind davon nur wenige Dokumente betroffen. Im Rahmen dieser Untersuchung werden diese den nutzergenerierten Inhalten zugeordnet.

Die Zuordnung der Ergebnisse zu den Ergebnistypen erfolgt durch zwei voneinander unabhängige Jurorinnen. Zur Überprüfung der Übereinstimmung werden zunächst die Ergebnisse von sechs zufällig ausgewählten Suchanfragen von beiden unabhängig voneinander zugeordnet. Von diesen 124 Ergebnissen werden 111 identisch beurteilt (89,52 %). Die Berechnung der Interrater-Reliabilität ergibt einen guten Wert (Cohen's Kappa = 0,83).

### Generierung der Testumgebung

Die Testumgebung besteht aus Webseiten, die die Suchergebnislisten nachbilden, und ist über einen Server der Universität Hildesheim verfügbar. Die Darstellung der Ergebnisse ist reduziert und enthält für jeden Treffer den Titel, die Kurzbeschreibung sowie die Anzeige-URL (Abb. 1). Zur Ergänzung werden darüber hinaus die zugehörige Suchanfrage, das Informationsbedürfnis und die Relevanzkriterien angezeigt. Die Ergebnisdokumente sind als JPEG-Dateien bzw. in Sonderformaten (PDF, DOC, PPT) hinterlegt.

Recht herzlichen Dank für Ihren lieben Brief und für Ihr Interesse an der schwierigen Thematik der Legasthenie. Auch Erwachsene stellt das ...  
[www.legasthen.de/leserbriefe/erwachs.htm](http://www.legasthen.de/leserbriefe/erwachs.htm)

Die Datenerfassung erfolgt im Test papierbasiert. Zum Einstieg werden soziodemografische Angaben erhoben. Die Testpersonen werden anschließend gebeten, ihren Kenntnisstand zur vorgegebenen Suchanfrage einzuschätzen, bevor sie die Kurzbeschreibungen der Top-20 Ergebnisse anhand eines Ausdruckes bewerten. Erst anschließend dürfen sie die tatsächlichen Ergebnisdokumente im Browser beurteilen. Abschließend bestimmen sie die Bedeutung der Qualitätskriterien für die vorliegende Suchanfrage und geben Auskunft darüber, ob sie ihr Informationsbedürfnis befriedigen konnten. Abbildung 2 verdeutlicht den Testablauf.

[illegible]

Zur Erfassung von Informationsqualität und Relevanz kommen verschiedene Maßeinheiten zum Einsatz. Relevanz ist das zentrale Konzept bei der Evaluierung von Retrievalsystemen. Daher werden auch in diesem Test die Relevanz der Ergebnisse und die der Trefferbeschreibungen beurteilt. Die Berechnung von etablierten Retrievalmaßen, beispielsweise der Precision, setzt ein binäres Relevanzurteil (relevant - nicht relevant) voraus. Um der



Hypertextstruktur des Internets besser gerecht zu werden, wird eine dritte Bewertungsmöglichkeit „verweist auf relevante Seite(n)“ angeboten. Dadurch können Webseiten berücksichtigt werden, die den Nutzer lediglich durch relevante Hyperlinks bei der Befriedigung von Informationsbedürfnisse unterstützen.

Die Bewertung der Trefferbeschreibung erfolgt dagegen auf einer binären Basis mit den Optionen „scheint relevant zu sein / würde ich anklicken“ und „scheint nicht relevant zu sein / würde ich nicht anklicken“. Eine weitere Differenzierung wird nicht verwendet, da den Testpersonen lediglich die Trefferbeschreibungen gezeigt werden.

Bei diesem Retrievaltest werden außerdem Urteile zur Vertrauenswürdigkeit, Verständlichkeit und Aktualität der Suchergebnisse erhoben. Da es den Testpersonen nicht möglich ist zu browsen, sind sie auf das unmittelbare Informationsangebot des Suchtreffers angewiesen. Die Beurteilung der Qualitätsdimensionen (nicht vertrauenswürdig – vertrauenswürdig, schlecht verständlich – gut verständlich, nicht aktuell – aktuell) erfolgt mit einer 5er-Skala. Für das Kriterium Aktualität wird zusätzlich die Option „kann ich nicht beurteilen“ angeboten. Zusammengefasst ergeben sich für jeden Treffer vier Bewertungsaufgaben. Abbildung 3 zeigt deren Darstellung im Retrievaltest.

<input type="checkbox"/> relevant	<input type="checkbox"/> nicht relevant	Zusatzinformation: <input type="checkbox"/> verweist auf relevante Seite(n)				
nicht vertrauenswürdig	<input type="checkbox"/> -2	<input type="checkbox"/> -1	<input type="checkbox"/> 0	<input type="checkbox"/> +1	<input type="checkbox"/> +2	vertrauenswürdig
schlecht verständlich	<input type="checkbox"/> -2	<input type="checkbox"/> -1	<input type="checkbox"/> 0	<input type="checkbox"/> +1	<input type="checkbox"/> +2	gut verständlich
nicht aktuell	<input type="checkbox"/> -2	<input type="checkbox"/> -1	<input type="checkbox"/> 0	<input type="checkbox"/> +1	<input type="checkbox"/> +2	aktuell
						<input type="checkbox"/> kann ich nicht beurteilen

Abbildung 3: Das Bewertungsschema für Ergebnisdokumente

Um das Testdesign zu überprüfen, werden sieben Pretests durchgeführt. Die Durchführung dauert zwischen 40 und 60 Minuten. Die papierbasierte Beurteilung der Trefferbeschreibungen wird von den Testpersonen positiv wahrgenommen. Sie scheint geeignet, die Arbeitsschritte gegeneinander abzugrenzen und erleichtert den Testpersonen, sich bei der Beurteilung der Ergebnisdokumente nicht auf die zuvor durchgeführte Beurteilung der Trefferbeschreibungen zu beziehen. Die Relevanzbeurteilung verläuft völ-

lig unproblematisch. Die Aufgabenstellung und der Ablauf sind gut verständlich und klar. Die benötigte Zeit und Belastung werden als akzeptabel beschrieben. Da einige Teilnehmer Schwierigkeiten haben, die Aktualität der Ergebnisdokumente zu beurteilen, wird die Option „kann ich nicht beurteilen“ ergänzt.

Die Retrievaltests finden zwischen dem 23.05.2011 und dem 09.06.2011 in den Computerräumen der Universität Hildesheim statt. Zur Testgruppe gehören 40 Testpersonen. Allen wird jeweils eine Suchanfrage zugeordnet. Zur Testgruppe zählen 30 Frauen und 10 Männer, die ausnahmslos Studierende pädagogischer Studiengänge an der Universität Hildesheim sind. An der Universität Hildesheim sind circa 80% der Studierenden Frauen. Der Frauenanteil liegt in den pädagogischen Studiengängen traditionell deutlich über dem Durchschnitt. 80% sind zwischen 21 und 25 Jahren alt. Ungefähr zwei Drittel der Teilnehmer befinden sich im Bachelorstudium und circa ein Drittel im Masterstudium. Mit einer Ausnahme bezeichnen sich alle als Experten und erfahrene Internetnutzer. Mehr als die Hälfte der Teilnehmer (55%) sucht mehrmals pro Woche mit Hilfe von Suchmaschinen nach fachbezogenen Informationen. 40% geben an, dies täglich zu tun.

Bei der Testdurchführung treten keine größeren Probleme auf. Die kognitive Belastung beeinflusst die Testergebnisse nicht negativ. Die Schwierigkeit der Bearbeitung der Aufgaben des Retrievaltests wird als sehr gut bewertet, was sich auch in der Beurteilung der Stimmung widerspiegelt. Bei 77,5% der Teilnehmenden zeigt sich keine Veränderung.

#### **4. Auswertung der Relevanz- und Qualitätsurteile der Nutzer**

In diesem Abschnitt werden einige der Ergebnisse dargestellt. Zunächst wird ein Überblick über Verteilung und Sichtbarkeit der Ergebnistypen gegeben. Anschließend werden Relevanz und Qualitätsurteile analysiert.

##### ***Sichtbarkeit***

Nahezu die Hälfte aller Treffer (48,8%) werden dem Ergebnistyp *Fachinformation* zugeordnet. 30,5% sind *professionell erstellte Inhalte* und 20,7% *nutzergenerierte Inhalte*. Damit sind fast 80% der Top-20 Treffer ein Produkt professioneller Contenterstellung. Abbildung 4 gibt einen Überblick über die Verteilung der Ergebnistypen unter den Top-20 Treffern.

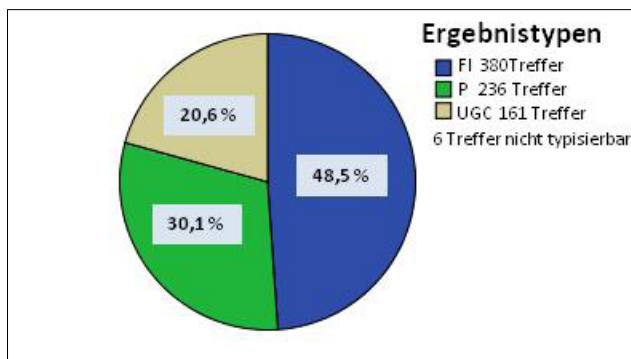


Abbildung 4: Verteilung der Ergebnistypen, N = 777

Die Ergebnistypen verteilen sich in unterschiedlichem Umfang über die Rankingpositionen der Suchergebnislisten. Dies ist insbesondere im Hinblick auf das Nutzerverhalten interessant, da die Nutzer häufig nur die ersten Treffer, bzw. die erste Seite der Suchergebnisliste betrachten. Auf der ersten Position sind *nutzergenerierte Inhalte* signifikant überrepräsentiert ( $p < 0,001$ ). Dies ist vornehmlich auf Dokumente der Wikipedia zurückzuführen. Auf den Positionen drei und vier sind Dokumente des Ergebnistyps *professionell erstellte Inhalte* signifikant überrepräsentiert ( $p < 0,05$ ). Abbildung 5 zeigt die Verteilung der Ergebnistypen über die Rankingpositionen. Das Ergebnis ist bemerkenswert, wenn man bedenkt, dass die ersten Ergebnisse die höchste Aufmerksamkeit bekommen.

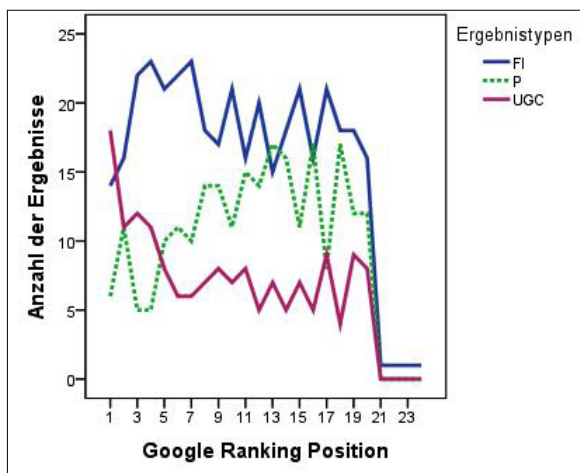


Abbildung 5: Verteilung der Ergebnistypen über die Rankingpositionen

## Relevanz- und Qualitätsurteile

Betrachtet man zunächst die Suchanfragen als Ganzes (Makroebene), so liefern alle untersuchten Suchanfragen mindestens vier relevante Treffer, d.h. die Ergebnisse befriedigen die fachspezifischen Informationsbedürfnisse. Auf die Frage, ob sie mit den Suchergebnissen ihr Informationsbedürfnis befriedigen können, antworten 74,4% mit „Ja“. Dies bedeutet im Umkehrschluss, dass 25,6% ihr Informationsbedürfnis nicht befriedigen können, obwohl es mindestens vier relevante Treffer gibt. Nach Broder (2002) sind bei informationsorientierten Suchanfragen mehrere Dokumente nötig, um ein Informationsbedürfnis zu befriedigen.

Zur differenzierten Analyse wird auf Ebene der Suchergebnisse (Mikroebene) die Precision genauer betrachtet. Tabelle 2 gibt einen Überblick über den Anteil relevanter Kurzbeschreibungen und Ergebnisdokumente am Ergebniskorpus.

	UGC	FI	P
Kurzbeschreibungen (Snippets)*	45,6%	57,0%	35,6%
Treffer**	52,8%	56,6%	41,9%

\*n = 775, \*\*n = 777

Tabelle 2: Der Anteil relevanter Ergebnisse in den verschiedenen Ergebniskategorien

Treffer des Typs *Fachinformation* haben eine deutlich größere Chance angeklickt zu werden. Die Analyse bestätigt signifikante Zusammenhänge zwischen den Ergebnistypen und den Relevanzurteilen. Die Relevanzbeurteilung der Ergebnisdokumente ergibt 56,6% relevante *FI-Dokumente*, 52,8% relevante *UGC-Dokumente* und 41,9% relevante *P-Dokumente*. Somit weisen die Ergebnisse des Typs *Fachinformation* die größte Relevanzwahrscheinlichkeit bezüglich der Treffer auf. Interessanterweise sind *nutzergenerierte Inhalte* relevanter als auf den ersten Blick angenommen. *Professionell erstellte Inhalte* von Akteuren ohne Bildungsauftrag erreichen in beiden Bereichen nur den dritten Platz.

Für jedes Suchergebnis liegt außerdem eine Beurteilung der Qualitätsdimensionen Vertrauenswürdigkeit, Aktualität und Verständlichkeit vor. Die Juroren geben darüber hinaus an, wie sie die Bedeutung der Qualitätsdimensionen für die vorgegebene Suchanfrage einschätzen. Die Mehrheit der

Juroren erachtet die Kriterien als wichtig für die vorliegende Suchanfrage. Verständlichkeit und Vertrauenswürdig sind von großer Bedeutung für die Suchanfrage und Befriedigung der Informationsbedürfnisse, während Aktualität eher nachrangig von Bedeutung ist (Vgl. Tabelle 3). In Tabelle 4 werden die Qualitätsurteile auf die Ergebnistypen projiziert.

Bedeutung des Qualitätskriteriums	Vertrauenswürdigkeit	Verständlichkeit	Aktualität
hoch	75,0 %	77,5 %	45,0 %
mittel	20,0 %	17,5 %	42,5 %
niedrig	5,0 %	5,0 %	12,5 %

Tabelle 3: Beurteilung der Bedeutung der Qualitätskriterien

Qualitätsurteil	Nutzergenerierte Inhalte (UGC)	Fachinformation (FI)	Professionelle Inhalte (P)
vertrauenswürdig	31,7 %	69,8 %	51,1 %
nicht vertrauenswürdig	45,6 %	9,2 %	17,9 %
unbestimmt	22,8 %	21,0 %	31,1 %
verständlich	65,7 %	74,7 %	57,9 %
schlecht verständlich	10,1 %	4,5 %	14,6 %
unbestimmt	24,2 %	20,8 %	27,5 %
aktuell	45,4 %	49,5 %	41,9 %
nicht aktuell	10,0 %	9,5 %	6,8 %
unbestimmt	11,5 %	8,7 %	16,2 %
nicht beurteilbar	33,0 %	32,3 %	35,0 %

Tabelle 4: Verteilung der Qualitätsurteile innerhalb der Ergebnistypen

Die Auswertung der Qualitätsurteile zeigt, dass lediglich 9,2% der *Fachinformationen* eine negative Beurteilung der Vertrauenswürdigkeit erhalten. Dem gegenüber werden 17,9% der *P-Treffer* und 45,6% der *UGC-Treffer* negativ bezüglich ihrer Vertrauenswürdigkeit beurteilt. *Fachinformationen* weisen den größten Anteil verständlicher Ergebnisse auf (74,7%), gefolgt von den *nutzergenerierten Inhalten* (65,7%) und *professionell erstellten Inhalten* (57,9%).

## 5. Ausblick

Untersuchungsgegenstand der in diesem Artikel vorgestellten Masterarbeit ist die Beurteilung bildungsbezogener Suchergebnisse von Universal-suchmaschinen. Die entwickelte Typologie der Suchergebnisse bietet interessante Einblicke jenseits von formalen oder technischen Aspekten. Die Kombination mit der Beurteilung der Relevanz und Qualität durch unabhängige Testpersonen bildet die Voraussetzung für die Analyse von Zusammenhängen.

Sowohl qualitativ als auch quantitativ überwiegen fachliche Suchergebnisse. Googles Universal-suche ist in der Lage, spezifische fachliche Informationsbedürfnisse zu befriedigen. Die Ergebnisse zeigen, dass nutzergenerierte Inhalte bei der Suche nach domänenspezifischen Informationen ebenfalls gut sichtbar sind und für die hier untersuchten Informationsbedürfnisse teilweise sogar besser sind, als Dokumente von professionellen Akteuren ohne besondere Expertise im Bildungssektor. Da Universalmaschinen vielfältige Informationsbedürfnisse bedienen, erscheint dies vertretbar im Sinne einer wünschenswerten Vielfalt von Informationsangeboten.

Mit der Identifikation von Ergebniskategorien, die durch ihre Charakteristik die Beurteilung von Suchergebnissen beeinflussen, wird ein Zusammenhang zwischen den Informationsproduzenten und der Evaluation sichtbar. Professionalität und Fachlichkeit der Informationsproduktion wirken sich positiv auf die Bewertung der Relevanz, der Vertrauenswürdigkeit und der Verständlichkeit aus. Die Analyse ergibt neue Erkenntnisse für den Anwendungsfall der horizontalen Suche zum Themenbereich Bildung.

Aus methodischer Sicht erscheint die Verknüpfung von etablierten Ansätzen zur Untersuchung der Retrievaleffektivität mit einer qualitativen Kategorisierung der Suchergebnisse vielversprechend. Aufgrund ihres explorativen Charakters sind die Ergebnisse der vorliegenden Arbeit jedoch nicht zu verallgemeinern. Die benutzerorientierte Art der Erschließung von Suchanfragen und Informationsbedürfnissen ist dennoch gut gelungen. Der in dieser Masterarbeit angewandte Ansatz bietet die Möglichkeit, neue und aussagekräftige Erkenntnisse zu generieren, die der bloßen Einteilung in relevante und nicht-relevante Ergebnisse neue Impulse geben können. Feinere Differenzierungen z.B. zwischen kollektiv und individuell generierten Inhalten könnten ebenfalls weitere Erkenntnisse liefern.

## Literaturverzeichnis

- Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, 36(2), 3–10.
- Baeza-Yates, R. (2009). User generated content: how good is it? In *Proceedings of the 3rd workshop on Information credibility on the web*, ACM 2009, New York, NY, USA, 1–2.
- Chai, K., Potdar, V., & Dillon, T. (2009). Content Quality Assessment Related Frameworks for Social Media. In O. Gervasi, D. Taniar, B. Murgente, A. Laganà, Y. Mun, & M. L. Gavrilova (Eds.), *Proceedings of the International Conference on Computational Science and Its Applications: Part II (ICCSA '09)* (pp. 791–805). Berlin, Heidelberg: Springer.
- Dickinger, A. (2011). The Trustworthiness of Online Channels for Experience- and Goal-Directed Search Tasks. In *Journal of Travel Research* 50, 4 (2011), 378–391.
- Gammer, O., Meißner, H., Preckel, M., & Oehlert, R. (2008). Vergleich der Relevanz von Treffern bei algorithmischen Suchmaschinen, Social Bookmarking-Seiten und Frage-Antwort-Diensten. In D. Lewandowski & C. Maaß (Eds.), *Web-2.0-Dienste als Ergänzung zu algorithmischen Suchmaschinen* (pp. 55–81). Berlin: Logos.
- Knight, S.-A., & Burn, J. (2005). Developing a Framework for Assessing Information Quality on the World Wide Web. *Information Science Journal*, 8, 159–172.
- Lewandowski, D. (2006). Themen und Typen der Suchanfragen an deutsche Web-Suchmaschinen. In F. Lehner, H. Nösekabel, & P. Keinschmidt (Eds.), *Multikonferenz Wirtschaftsinformatik 2006 (MKWI ,06). Proceedings 20. - 22. Februar 2006* (pp. 33–43). Berlin: Gito.
- Lewandowski, D., & Höchstätter, N. (2007). Qualitätsmessung bei Suchmaschinen. System- und nutzerbezogene Evaluationsmaße. *Informatik Spektrum*, 30(3), 1–11.
- Lewandowski, D. and Spree (2011). U. Ranking of Wikipedia articles in search engines revisited: Fair ranking for reasonable quality? In *Journal of the American Society for Information Science and Technology* 62, 1 (2011), 117–132.
- Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future re-

search. *Journal of the American Society for Information Science and Technology*, 58(13), 2078–2091.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, 53(2), 145–161.

- 1 [http://www.youtube.com/t/press\\_statistics](http://www.youtube.com/t/press_statistics)
- 2 <http://blog.flickr.net/de/2012/04/25/begrust-den-neuen-uploadr/>
- 3 [http://blog.nielsen.com/nielsenwire/online\\_mobile/buzz-in-the-blogsphere-millions-more-bloggers-and-blog-readers/](http://blog.nielsen.com/nielsenwire/online_mobile/buzz-in-the-blogsphere-millions-more-bloggers-and-blog-readers/)
- 4 [www.bildungsserver.de](http://www.bildungsserver.de)